個別銘柄情報と銘柄間情報を考慮した銘柄埋め込み手法の提案

高柳剛弘 [1]　坂地泰紀 [1]　和泉潔 [1]
[1] 東京大学　大学院工学研究科
m2021ttakayanagi@socsim.org

概要

本研究では個別銘柄と銘柄間関係の特徴を同時に学習する新たな銘柄埋め込み手法を提案する．提案手法では，因果チェーンにより銘柄間ネットワークを構築し，アニュアルレポートのテキストデータから BERT を用いて個別銘柄独自の特徴を学習するのに加えて，銘柄間ネットワークに対して GCN を用いることで銘柄間のクロスセクショナルな関係を同時に学習することを可能にした．東京証券取引所の上場銘柄を対象に行った実験により，提案手法が既存手法の性能を大きく上回ることを確認し，銘柄独自の特徴と銘柄間のクロスセクショナルな関係を同時に学習することの有用性を示した．

1　はじめに

今日の金融市場において機械学習やテキストマイニングなどの技術の活用が進んでいる．金融市場や投資家にまつわる膨大なデータの蓄積やそれに伴う大規模データ解析技術の進展は金融分野における技術発展を加速させている．特に資産運用分野では株価や経済指標の数値予測 [1, 2, 3]，銘柄の推薦 [4] や投資家のサポート [5, 6] などデータを駆使した多様なアプリケーションが生まれた．これらのアプリケーションには離散オブジェクトである個別銘柄やファンドをベクトルで表現するアプローチが多く用いられ [1, 2, 3, 4, 5, 6]，個別銘柄やファンドのベクトル埋め込みは資産運用分野における技術発展の基盤となっている．

一般に離散オブジェクトをベクトルにより表現する際にはそのオブジェクトの特徴を捉える必要がある．特に銘柄の特徴は銘柄自身の情報と銘柄間のクロスセクショナルな関係の情報の二つにより表すことができると考えられる．銘柄自身の特徴を表現するためのデータとして企業の株価や財務情報など伝統的なデータのみならず，テキストデータなどのオルタナティブデータの利活用が進んでいる．[5] は銘柄埋め込みを作成する際のテキストデータの有用性について示した．

一方で銘柄間関係の表現については，企業のサプライチェーン，取引構造や株式所有構造などの企業間の関係を模した企業間ネットワークを利用する手法が多く用いられている [1, 2, 3, 4, 6]．このように銘柄のベクトル埋め込みには二つのアプローチが存在し，個別銘柄の情報，銘柄間の情報共に有用であることが示されている．

しかしながら既存研究では銘柄自身の情報と銘柄間関係の情報の双方を同時に考慮することができておらず，銘柄の一側面のみに注目した研究になっている．加えて，企業間の関係はサプライチェーンや取引データなど方向性のある関係である．このことから有向グラフにより企業間の関係を抽出する方法が妥当であると考えられるが，既存研究では無向グラフにより企業ネットワークが構築され，有向グラフの有用性は検証されていない．

これらの課題を解決するために，本研究では BERT と GCN（Graph Convolutional Network）を組み合わせることで個別銘柄の情報と銘柄間関係を同時に学習させる手法を提案する．提案手法では初めに因果チェーンを用いて銘柄間ネットワークを構築する．次に銘柄のテキスト情報に対して BERT を用いることで得られる個別銘柄独自の特徴をノード特徴量として用いる．最後に銘柄間ネットワークに対して GCN を用いることで銘柄とその周辺銘柄の関係性を考慮した銘柄埋め込みを獲得する．

実験を通して提案手法は先行研究の性能を上回り，銘柄埋め込みに対して個別銘柄と銘柄間の関係性を同時に学習させることの有用性が示された．加えて，銘柄を因果チェーンで結んだ有向グラフを作成することで，企業間の関係性を示すために有向グラフが有用である可能性を示した．

2 先行研究

銘柄のベクトル埋め込みの研究は大きく分けて二つに分けることができる．一つ目は個別銘柄の特徴に注目するアプローチである．このアプローチでは銘柄の産業区分や事業の内容など銘柄独自の情報をベクトル表現に織り込むことにより銘柄のベクトル表現を獲得している．[5] の研究ではアニュアルレポートのテキスト情報から個別企業の特徴を抽出することで銘柄のベクトル表現を獲得した．二つ目は銘柄間のクロスセクショナルな関係性に注目するアプローチである．このアプローチでは企業のサプライチェーン，取引関係や株式所有構造などの銘柄間の関係の情報をベクトル表現に織り込むことにより企業のベクトル表現を獲得している [1, 2, 3, 4, 6].

[1] では株主所有構造に [3] ではニュースにおける銘柄の共起性に [6] では銘柄と投資信託の内包関係によりネットワークを構築し Node2Vec を用いて埋め込みを獲得している．[2] では株式所有構造，企業の産業区分，ニュースのトピック情報に，[4] では銘柄の産業区分によりネットワークを作成し GCN を用いて埋め込みを獲得している．

3 提案手法

本研究の目的はターゲット銘柄 T_i についての d 次元のベクトル $h_i \in \mathbb{R}^d$ を得ることである．提案手法は図 1 の通りである．初めに因果チェーンからターゲット銘柄 T_i を中心とするサブグラフを抽出する（STEP1）．次に各ノードのノード特徴量を BERT に入力する（STEP2）．最後に GCN を用いてノード特徴量を更新する（STEP3）．

3.1 BERT

自然言語処理のタスクで優れた性能を示している BERT[7] を用いてテキストデータの埋め込み表現を獲得する．BERT は東北大学の乾研究室[1)]が公開している Pretrained Japanese BERT model を利用した．BERT は専門単語が多い領域である金融分野において金融コーパスを用いて再事前学習（further pretraining）をすることでの精度向上が報告されている [8]．これに従い本研究では金融コーパスを用いた再事前学習を行い，再事前学習された BERT モデルを用いて学習を行った．

$$h_{BERT} = BERT(x_i) \tag{1}$$

ここで x_i はターゲット銘柄 i のテキストデータであり，$h_{BERT} \in \mathbb{R}^e$ は BERT の出力である．

3.2 GCN, GAT

GCN (Graph Convolutional Network)[9] は各ノードが隣接ノードから情報を集約（aggregate）し，自身の情報を更新（update）していくニューラルネットワークの手法である．

$$h_{GCN} = \sigma\left(\sum_{j \in N_i} \hat{D}^{-\frac{1}{2}} \hat{A} \hat{D}^{-\frac{1}{2}} h_j W + b\right) \tag{2}$$

ここで，σ は $ReLU$ 関数などの非線形活性化関数，N_i はターゲット銘柄 i の隣接ノード，$\hat{A}$ は グラフの隣接行列 A + 単位行列 I_N，$\hat{D}$ は $\sum_j \hat{A}_{ii}$，h は GCN の入力，W は重み行列，b はバイアスを示している．

GAT (Graph Attention Network)[10] は隣接ノードの情報を集約（aggregate）する際にアテンションメカニズムを用いることにより，隣接ノードの重要性を考慮に入れるモデルである．GAT のアテンション係数は以下のように計算される．

$$a_{ij} = \frac{exp\left(LeakyReLU\left(a\left[Wh_i \| Wh_j\right]\right)\right)}{\sum_{k \in N_i} exp\left(LeakyReLU\left(a\left[Wh_i \| Wh_k\right]\right)\right)} \tag{3}$$

ここで, a は attention mechanism（ここでは一層の全結合層）であり，$\|$ は連結演算（Concatenation operation）を示し，活性化関数には $negativeslope = 0.2$ の LeakyReLU を適応している．

3.3 残差接続

残差結合は GCN（または GAT）の出力に，BERT の出力をそのまま足しこむことで，BERT により得られた銘柄独自の情報を「残す」役割をはたす．残差結合は以下のように計算される．

$$\begin{aligned} h_{BERT} &= BERT(x_i) \\ h_{GCN} &= GCN(h_{BERT}) \\ h_{res} &= h_{BERT} + h_{GCN} \end{aligned} \tag{4}$$

3.4 損失関数

本研究では TOPIX-17 と TOPIX-33 のラベルを予測するマルチタスク学習を行う．損失関数は TOPIX-17 と TOPIX-33 の損失の和で示される．

$$Loss = Loss^{TOPIX-17} + Loss^{TOPIX-33} \tag{5}$$

1） https://github.com/cl-tohoku/bert-japanese

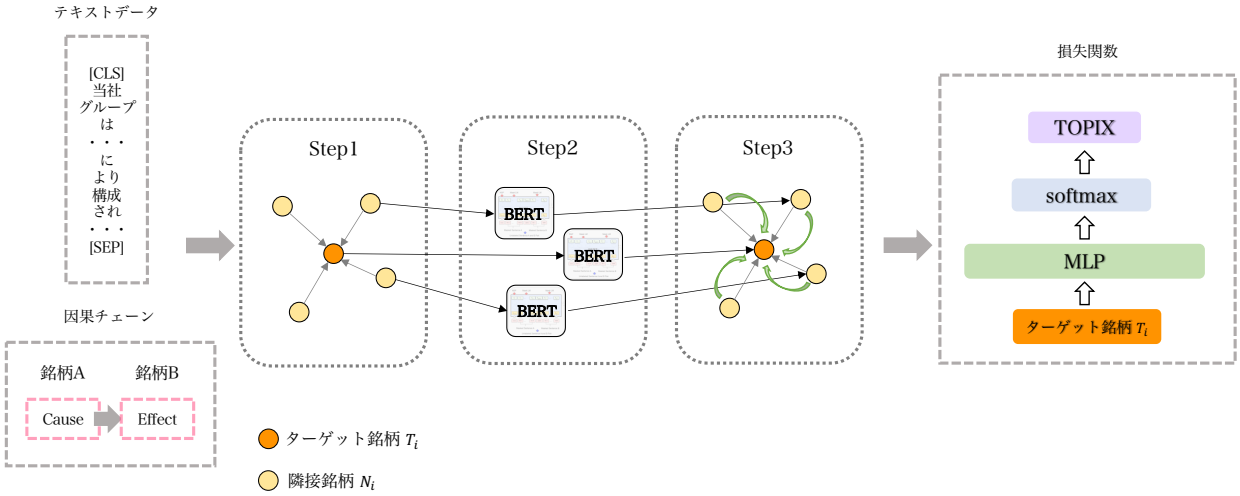

図 1　提案手法の概要

TOPIX-17，TOPIX-33 の損失はそれぞれ以下のように計算される．

$$y_i^{TOPIX} = Softmax\left(W^{TOPIX} h_i + b^{TOPIX}\right)$$
$$Loss^{TOPIX} = \sum_{i \in \Omega} CE\left(d_i^{TOPIX}, y_i^{TOPIX}\right) \quad (6)$$

ここで h_i は GCN の出力や残差接続された出力である．Ω は全銘柄を示し，CE は交差エントロピー損失関数を示す．ここでは TOPIX-17 と TOPIX-33 に関して同様の計算を行うため，合わせて TOPIX と記した．

4　実験設定

本研究では TOPIX-17 と TOPIX-33 のラベルによる類似企業抽出を用いた．ベースラインモデル [5] との性能を比較する実験（第 1 実験）に加えて，有向グラフ，無向グラフそれぞれの性能の比較実験（第 2 実験）を行った．

4.1　データセット

因果チェーン　本研究では因果チェーンを用いて銘柄間のネットワークを作成する．因果チェーン [11] は因果関係について記されたテキストデータを解析することで，人間が認知するような因果関係を抽出したデータである．[12] では因果チェーンがサプライチェーンや取引関係のような企業間の関係を表していることが示唆された．

ここでは因果チェーンの構築手法を [11] に基づき概観する．因果チェーンは以下の 3 ステップにより構築される．

Step1　SVM を使用して日本の決算短信の要約から因果関係の表現を含む文を抽出．

Step2　原因と結果を示す構文パターンを定義し，そこに Step1 で抽出した文における因果関係を抽出．

Step3　抽出された因果関係について，その終端ノード (結果側) の表現と一定の類似性がある他の決算短 信の起点ノード (原因側) とを連結し因果チェー ンを構築する．

また，因果関係にあるそれぞれの決算短信の発行日と企業名をキーとして，因果の出現数を集計する．本研究ではこの因果チェーンのデータを用いて企業間の有向グラフを作成する．ここで，因果の出現回数が 100 以下のエッジは偶然の弱い繋がりと考えてグラフから捨象している．

テキストデータ　本研究では決算短信，有価証券報告書やアニュアルレポートなど日本の金融レポートを集めた CoARiJ[2] データセットを用いた．再事前学習の際には 2014 年から 2018 年までのドキュメント全体を用い，学習時には 2018 年の各銘柄における「事業の概要」のテキストデータを各銘柄のノード特徴量として利用した．

4.2　学習

再事前学習 (further pretraining)　CoARiJ 全体の 207332 個のテキストファイルに対して Masked language model と Next sentence prediction により 20 万ステップに渡り BERT の全ての層に対する再事前学

2）https://github.com/chakki-works/CoARiJ

表 1 第 1 実験の結果

	TOPIX-17			TOPIX-33		
	MAP@5	MAP@10	MAP@50	MAP@5	MAP@10	MAP@50
BERT	0.634	0.591	0.504	0.555	0.526	0.461
BERT → GCN（GCN のみ学習）	0.49	0.399	0.197	0.414	0.32	0.149
BERT+GCN	0.701	0.666	0.548	0.585	0.538	0.427
BERT+GCN+Residual connection	**0.772**	**0.741**	**0.646**	**0.665**	**0.616**	**0.535**
BERT+GAT	0.586	0.526	0.368	0.471	0.42	0.301
BERT+GAT+Residual connection	0.739	0.705	0.575	0.631	0.588	0.479

表 2 第 2 実験の結果

		TOPIX-17			TOPIX-33		
		MAP@5	MAP@10	MAP@50	MAP@5	MAP@10	MAP@50
Directed	**BERT+GCN+Residual connection**	**0.772**	**0.741**	**0.646**	**0.665**	**0.616**	**0.535**
	BERT+GAT+Residual connection	**0.739**	**0.705**	**0.575**	**0.631**	**0.588**	**0.479**
Undirected	BERT+GCN+Residual connection	0.689	0.653	0.552	0.564	0.518	0.449
	BERT+GAT+Residual connection	0.732	0.704	0.598	0.594	0.552	0.459

習（further pretraining）を行った．

学習 本研究では 3016 銘柄を対象に実験を行い，訓練データ，検証データ，テストデータを 2200 銘柄, 316 銘柄, 500 銘柄と分割した．各銘柄のノード特徴量には「事業の概要」の 512 トークンを用いた．ターゲット銘柄を中心としたサブグラフごとに学習を行うが，サブグラフ全体で勾配の逆伝播を行うとデータリーケージの問題が発生するので，ターゲット銘柄の勾配のみを逆伝播させる．BERT は最終層のみ学習し，エポック数は 30，学習率は 0.001，オプティマイザは Adam を用いた．

4.3 評価指標

本研究では TOPIX17，TOPIX33 のラベルを用いた類似度企業抽出を評価指標に設定した．[5, 13] に基づき，Mean Average Precision at K (MAP@K) を用いて K = 5, 10, 50 でそれぞれ実験を行った．評価の手順として，まずはそれぞれの銘柄に対するコサイン類似度が最も大きい K 個の銘柄を選定し，それらを TOPIX17, TOPIX33 をもとに MAP@K を用いて評価した．

5 結果

提案手法の性能を確かめるために，ベースラインである [5] の研究や提案手法内でのさまざまなモデルを組み合わせた実験を行い，提案手法が既存手法をアウトパフォームしていることを確認した (表 1)．ここで BERT 単体は個別銘柄独自の特徴のみ学習し，BERT→GCN（GCN のみ学習）では銘柄間関係をメインに学習し，そして BERT+GCN では個別銘柄独自の特徴と銘柄間関係を同時に学習している．結果から銘柄の埋め込みに対して個別銘柄独自の情報とクロスセクショナルな情報を同時に学習させることの有用性が示唆された．さらに，BERT+GCN よりも BERT+GCN+Residual connection で精度が高いという結果は，残差結合により個別銘柄独自の特徴を「残す」ことで精度が向上したと解釈できる．

表 2 は因果チェーンから有向グラフと無向グラフを作成し精度を比較した実験の結果である．BERT+GCN においても BERT+GAT においても有向グラフによる学習の精度がより高いことから，企業間の関係性を表現するのに有向グラフを用いることの有用性を示すことができた．サプライチェーンや取引関係など実際の企業関係のおおくは方向性のある関係であるために，有向グラフを用いることで性能が向上したと考えられる．

6 終わりに

本研究では，BERT と GCN を組み合わせることで銘柄自身の情報と銘柄間のクロスセクショナルな関係を同時に織り込んだ新たな銘柄埋め込み手法を提案した．

実験結果により提案手法は [5] の性能を上回り最高精度を達成することを確認した．加えて，因果チェーンを用いて有向グラフを作成することで，銘柄埋め込みに対する有向グラフの有用性を示した．

謝辞

本研究は大和証券グループの支援を受けたものである．加えて，本研究は JSPS 科研費 JP21K12010 と JST 未来社会創造事業 JPMJMI20B1 の助成を受けたものである．

参考文献

[1] Yingmei Chen and Zhongyu Wei. Incorporating corporation relationship via graph convolutional neural networks for stock price prediction. pp. 1655–1658. Association for Computing Machinery, 2018.

[2] Jiexia Ye, Juanjuan Zhao, Kejiang Ye, and Chengzhong Xu. Multi-graph convolutional network for relationship-driven stock movement prediction. pp. 6702–6709, 2021.

[3] Qiong Wu, Christopher G Brinton, Zheng Zhang, Andrea Pizzoferrato, Zhenming Liu, and Mihai Cucuringu. Equity2vec: End-to-end deep learning framework for cross-sectional asset pricing ; equity2vec: End-to-end deep learning framework for cross-sectional asset pricing. 2021.

[4] Jianliang Gao, Xiaoting Ying, Cong Xu, Jianxin Wang, Shichao Zhang, and Zhao Li. Graph-based stock recommendation by time-aware relational attention network. **ACM Transactions on Knowledge Discovery from Data**, Vol. 16, , 2021.

[5] Tomoki Ito, Jose Camacho Collados, Hiroki Sakaji, and Steven Schockaert. Learning company embeddings from annual reports for fine-grained industry characterization. **Proceedings of the Second Workshop o Financial Technology and Natural Language Processing**, pp. 27–33, 2020.

[6] Vipul Satone, Dhruv Desai, and Dhagash Mehta. Fund2vec: Mutual funds similarity using graph learning. **arXiv preprint arXiv:2106.12987**, 6 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, 2019.

[8] Yuta Niki, Hiroki Sakaji, Kiyoshi Izumi, and Hiroyasu Matsushima. 再事前学習した bert を用いた金融文書中の因果関係知識有無の判別. pp. 3Rin439–3Rin439, 2020.

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2017.

[10] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. 2018.

[11] Izumi Kiyoshi and Sakaji Hiroki. Economic causal-chain search using text mining technology. pp. 23–35, 2019.

[12] Kei Nakagawa, Shingo Sashida, Hiroki Sakaji, and Kiyoshi Izumi. 経済因果チェーンを用いたリードラグ効果の実証分析. 2019.

[13] Kuifei Yu, Baoxian Zhang, Hengshu Zhu, Huanhuan Cao, and Jilei Tian. Towards personalized context-awarerecommendation by mining context logsthrough topic models. 2012.